John von Neumann Institute for Computing

The Two Aspects of the Protein Folding Problem

Harold A. Scheraga, Adam Liwo, Stanislaw Ołdziej,
Cezary Czaplewski, Mey Khalili, Jorge A. Vila,
Daniel R. Ripoll

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 37-44 , 2006.

http://www.fz-juelich.de/nic-series/volume34

The Two Aspects of the Protein Folding Problem

Harold A. Scheraga[1], Adam Liwo[1,2], Stanislaw Ołdziej[1,2], Cezary Czaplewski[1,2], Mey Khalili[1], Jorge A. Vila[1,3], and Daniel R. Ripoll[4]

[1] Baker Laboratory of Chemistry and Chemical Biology,
Cornell University, Ithaca, New York 14853-1301
E-mail: has5@cornell.edu

[2] Faculty of Chemistry, University of Gdansk,
80-952 Gdansk, Poland

[3] Universidad Nacional de San Luis, IMASL-CONICET,
Ejército de los Andes 950, (5700) San Luis, Argentina

[4] Computational Biology Service Unit-Cornell Theory Center,
Cornell University, Ithaca, New York 14853

A physics-based approach to the protein folding problem is presented. It is concerned with the computation of folding pathways and final native structures, given the amino acid sequences, an empirical all-atom potential energy function, and a procedure to identify the global minimum of the potential energy. Whereas the all-atom approach has provided three-dimensional structures of relatively small molecules and for helical proteins containing up to 46 residues, it has been necessary to develop a hierarchical approach to treat larger proteins. In the hierarchical approach, global optimization was originally carried out with a simplified united residue (UNRES) description of a polypeptide chain to locate the *region* in which the global minimum lies. Conversion of the UNRES structures in this region to all-atom structures is followed by a local search in this region. The performance of this physics-based approach in successive CASP blind tests for predicting protein structure is described. More recently, a molecular dynamics treatment with UNRES has been introduced to compute not only native structures but also folding pathways.

1 Introduction

Ever since Anfinsen[1] demonstrated that a polypeptide chain can fold spontaneously into the three-dimensional structure of a native protein, experimental and theoretical chemists have tried to determine the interactions that govern the folding process (the protein-folding problem). Actually, there are two protein-folding problems: (a) determination of the folding pathways, and (b) identification of the folded native structure corresponding to the global minimum of the potential energy according to Anfinsen's thermodynamic hypothesis[1] that the native structure is the thermodynamically most stable one.

Early experimental efforts[2] were devoted to identifying inter-residue interactions in the native structure, and subsequent experiments identified multiple folding pathways[3,4]. Early theoretical approaches were based on the use of empirical interatomic potential energy functions[5] together with a menu of procedures[6] for global optimization of the potential energy. Some of these early efforts are summarized in reference 7. This article is devoted to recent theoretical work on the two protein-folding problems.

While the use of all-atom potential energy functions led to success for relatively small proteins, containing up to 20 residues[7], it was only after the development of large-scale

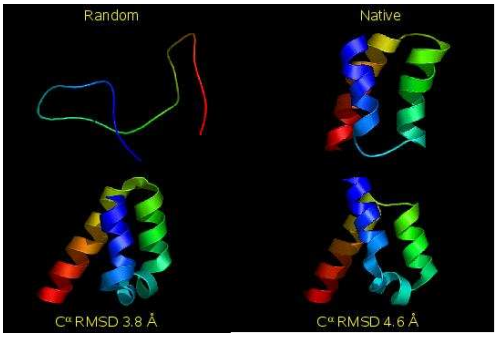

Figure 1. Results of all-atom calculations on protein A[9], starting from a randomly-generated structure. The native structure and the lowest-energy structures obtained with two different hydration models[12,13] area also illustrated.

parallel-processor computers[8] that it has been possible to carry out computations with all-atom potentials for larger proteins, the largest one treated thus far being the 46-residue staphylococcal protein A[9].

2 All-Atom Calculations on Protein A

Calculations on protein A[9] were carried out with the empirical potential function ECEPP/3[10] and the electrostatically-driven Monte Carlo (EDMC) procedure[11], together with two implicit hydration models, OONS[12] and SRFOPT[13], starting from four different random conformations, one of which is illustrated schematically[9] in Figure 1. Three of the four runs converged to the same native-like fold illustrated in Figure 1 for each of the two hydration models; the fourth converged to the mirror-image conformation.

Protein A is larger than the 36-residue α-helical protein from the villin headpiece, for which all-atom simulations, starting from an extended structure, were previously carried out by other groups[14,15]. Those simulations were carried out with explicit solvent, which increases the computing time considerably compared to the time required for the implicit solvent models used in our simulations[9].

It is not yet clear what the largest size protein is that can be treated by our all-atom EDMC procedure. Nevertheless, it is encouraging that an all-atom representation of the chain, and global optimization of the corresponding potential energy, can identify the native-like fold without resorting to knowledge-based information in the search procedure.

3 Hierarchical Procedure to Predict Protein Structure

Without waiting for further extensions of the EDMC procedure, we have developed a hierarchical procedure to treat larger proteins containing both α and β folded portions. In this procedure, global optimization is carried out by using a Conformational Space Annealing (CSA) method[16,17] with a united-residue (UNRES) representation of the protein chain[18–20]. This is the key stage of the hierarchical algorithm. It is designed to locate the *region* of the global minimum rapidly and efficiently. The lowest-energy structures

obtained from the UNRES representation in this stage are then converted to the all-atom representation[21,22], and a local search is carried out in the restricted region located with the UNRES/CSA approach. This is accomplished with the EDMC method and the ECEPP/3 force field[10], together with the SRFOPT hydration model[13]. Initially, the backbone of the chain is constrained to the structures obtained by UNRES and CSA, but the constraints are gradually reduced as the calculations proceed.

The UNRES model[18–20] consists of a virtual-bond chain, i.e., a sequence of α-carbons, united peptide groups, and united side chains represented by ellipsoids whose size depends on the nature of the amino acid residue. The α-carbons are not centers of interaction, but merely serve to locate the backbone. The centers of interaction are the united peptide groups and united side chains, with a united-residue potential given as the sum of interactions involving side chain-side chain, side chain-peptide, peptide-peptide, virtual-bond torsional, virtual-bond double torsional, virtual-bond-angle bending, internal side-chain motional, and multi-body (correlation) energies. The variables to change conformation are the angles between virtual bonds, the torsional angle for rotation about the virtual bonds, and the position angle and rotational angle of the side chains.

The CSA method[16,17], used to search conformational space, starts with an initial set of widely-spaced UNRES minima. CSA is based essentially on a build-up and genetic algorithm to force these minima to coalesce to the *region* of the global minimum. All UNRES minimum-energy conformations in the final coalesced clusters are converted to the all-atom representation[21,22], and the global optimization search is continued from these starting conformations with the EDMC procedure[11].

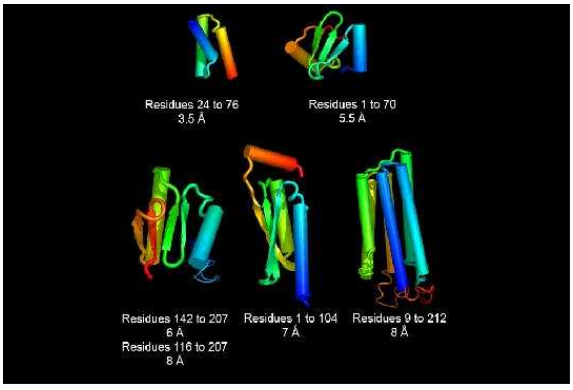

Figure 2. Results of blind CASP6 predictions.

4 CASP Results with Hierarchical Procedure

After carrying out initial tests of the hierarchical procedure on proteins of known structure, we participated in successive blind tests (CASP, Critical Assessment of Protein Structure Prediction), beginning with CASP3 in 1998. Various improvements of the procedure and

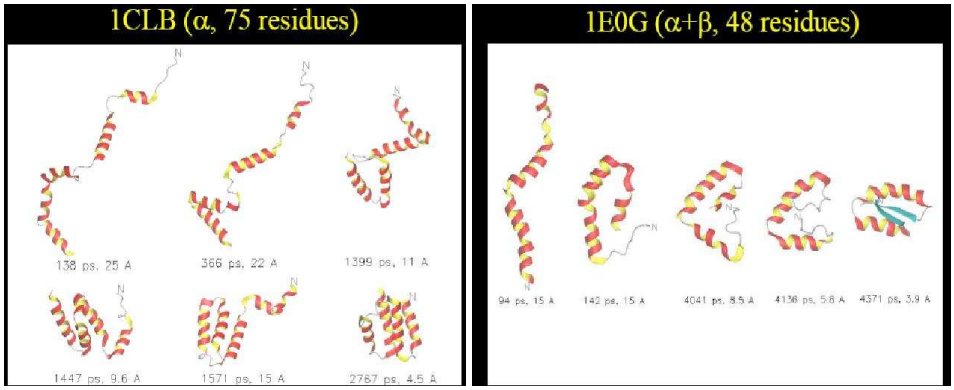

Figure 3. Right: Example of a fast-folding pathway of 1CLB obtained in Langevin dynamics simulations. The N-terminus of the chain is marked for tracing purposes. Left: Examples of a folding pathway of 1E0G obtained in Langevin dynamics simulations. The N-terminus of the chain is marked for tracing purposes.

its associated physics were implemented in successive tests. Some results from the most recent test (CASP6 in 2004) are illustrate in Figure 2.

5 Calculations of Folding Pathways

As pointed out in the Introduction (section 1), a second type of protein folding problem is the computation of the structural pathways by which the completely unfolded polypeptide chain proceeds to the folded native conformation, i.e., the progression from a given unfolded state (with no native contacts or native hydrogen bonds) to the final folded structure. One approach assumes that both the initial and final structures are known, and computes the folding pathways by used of the stochastic difference equation method of Elber *et al*[23], and has been applied with a full-atom treatment to protein A[24]. A second approach, applied to protein A and larger proteins, makes use of the UNRES force field, in which the fast degrees of freedom are averaged out, and carries out Langevin molecular dynamics[25–28]. The structure of protein A was obtained within an RMSD of 3.0Å. The largest protein to which th is second approach was applied is the 75-residue, α-helical protein 1CLB, and the structure was obtained within an RMSD of 4.5Å (see Figure 3(left)).

The structure of a 48-residue α + β protein, 1E0G, was obtained within an RMSD of 3.9Å (see Figure 3(right)). This molecular dynamics technique provides not only the final folded structure, but also the intermediate structures along the folding pathways. Further, by computing 400 trajectories to obtain good statistics, it was possible to compute the folding kinetics of protein A. As illustrated in Figure 4, the kinetics is either two-state (single exponential) or three-state, depending on the quantity that is computed (or measured experimentally), i.e., two-state for helix content (without considering interhelical interactions), or three-state (bi-exponential) if the RMSD is computed (which includes interhelical interactions).

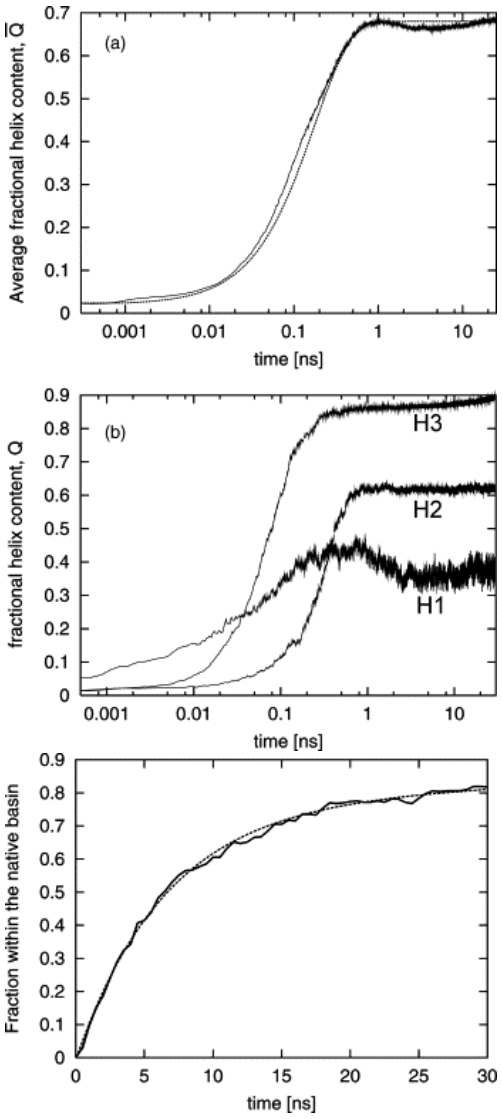

Figure 4. (a) Plot of the helix content of protein A ($\bar{Q}$), averaged over 400 trajectories (continuous line) and the exponential fit to $\bar{Q}$ (broken line). (b) Plots of the helix content of segments corresponding to helices H1, H2, and H3, from the N- to the C-terminus. (c) Plot of the fraction of native structure (secondary + tertiary) as a function of time (continuous line), and the fit of a bi-exponential function to these data (broken line).

6 Conclusions

The evolution of computational methodology has led from an all-atom treatment illustrated in Figure 1 to a hierarchical treatment of larger proteins (Figure 2) and to a longer-time molecular dynamics (MD) approach to treat not only the final folded structures but also

the intermediate structures and the folding kinetics illustrated in Figures 3, 4, respectively. Current work is focused on use of the UNRES/MD approach with improvement of the force field and inclusion of entropic effects.

Acknowledgements

This research was supported by grants from NIH (GM-14312, TW-7193, TW-6335) and NSF (MCB00-03722), and was carried out with resources of (a) our 392-double processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, (b) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, (c) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk, and (d) the Cornell Theory Center which receives funding from Cornell University, New York State, Federal agencies foundations, and corporate partners.

References

1. C. B. Anfinsen *Principles that Govern Folding of Protein Chains* , Science **181**, 223–230 (1973).
2. H. A. Scheraga *Structural studies of pancreatic ribonuclease*, Fed. Proc. **26**, 1380-1387 (1967).
3. T. E. Creighton and D. P. Goldberg *Kinetic role of a meta-stable native-like two-disulfide species in the folding transition of bovine pancreatic trypsin inhibitor*, J. Mol. Biol. **179**, 497-526 (1984).
4. D. M. Rothwarf, Y. J. Li and H. A. Scheraga *Regeneration of bovine pancreatic ribonuclease A. Detailed kinetic analysis of two independent folding pathways*, Biochemistry **37**, 3767-3776 (1998).
5. H. A. Scheraga *Calculations of conformations of polypeptides*, Adv. Phys. Org. Chem. **6**, 103-184 (1968).
6. J. A. Vila, H. A. Baldoni and H. A. Scheraga *Position dependence of the ^{13}C chemical shifts of α-helical model pptides. Fingerprint of the 20 naturally occurring amino acids*, Protein Science **13**, 2939-2948 (2004).
7. H. A. Scheraga, A. Liwo, S. Ołdziej, C Czaplewski, J. Pillardy, D.R. Ripoll, J.A. Vila, R. Kazmierkiewicz, J.A. Saunders, Y.A. Arnautova, A. Jagielski, M. Chinchio and M. Nanias *The protein folding problem: Global optimization of force fields*, Frontiers in Bioscience **9**, 3296-3323 (2004).
8. J. Lee, J. Pillardy, C. Czaplewski, Y. Arnautova, D. R. Ripoll, A. Liwo, K.D. Gibson, R.J. Wawak, and H. A. Scheraga *Efficient parallel algorithms in global optimization of potential energy functions*, Comput. Physics Commun. **128**, 399-411 (2000).
9. J. A. Vila, D. R. Ripoll and H. A. Scheraga *Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures*, Proc. Natl. Acad. Sci. U.S.A. **100**, 14812-14816 (2003).
10. G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey and H. A. Scheraga *Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm,*

with application to proline - containing peptides,
J. Phys. Chem. **96**, 6472-6484 (1992).
11. D.R. Ripoll, A. Liwo, and H. A. Scheraga *New Developments of the electrostatically driven Monte Carlo method: Test on the membrane-bound portion of melittin*,
Biopolymers **46**, 117-126 (1998).
12. T. Ooi, M. Oobatake, G. Némethy and H. A. Scheraga *Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides*,
Proc. Natl. Acad. Sci., U.S.A. **84**, 3086-3090 (1987). Erratum: ibid. **84**, 6015 (1987).
13. J. Vila, R. L. Williams, M. Vasquez and H. A. Scheraga *Empirical solvation models can be used to differentiate native from near - native conformations of bovine pancreatic trypsin inhibitor*,
Proteins: Structure, Function, and Genetics **10**, 199-218 (1991).
14. Y. Duan and P.A. Kollman *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution*,
Science **282**, 740-744 (1998).
15. B. Zagrovic, C.D. Snow, M.R. Shirts and V.J. Pande *Simulation of folding for a small α-helical protein in atomistic detail using worldwide-distributed computing* J. Mol. Biol. **323**, 927-937 (2002).
16. J. Lee, H. A. Scheraga and S. Rackovsky *New optimization method for conformational energy calculations on polypeptides: Conformational space annealing*,
J. Comput. Chem. **18**, 1222-1232 (1997).
17. J. Lee and H.A. Scheraga *Conformational space annealing by parallel computations: Extensive conformational search of Met-enkephalin and of the 20-residue membrane-bound portion of melittin*,
Intl. J. of Quantum Chem. **75**, 255-265 (1999).
18. A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky and H.A. Scheraga *A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data*,
J. Comput. Chem. **18**, 849-873 (1997).
19. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Oldziej and H.A. Scheraga *A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization*,
J. Comput. Chem. **18**, 874-887 (1997).
20. A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga *United-residue force field for off-lattice protein-structure simulations; III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials*,
J. Comput. Chem. **19**, 259-276 (1998).
21. R. Kazmierkiewicz, A. Liwo and H.A. Scheraga *Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte-Carlo method*,
J. Comput. Chem. **23**, 715-723 (2002).
22. R. Kazmierkiewicz, A. Liwo and H.A. Scheraga *Addition of side chains to a known backbone with defined side-chain centroids*,
Biophys. Chem. **100**, 261-280 (2003). Erratum: Biophys. Chem. **106**, 91 (2003).

23. R. Elber, A. Ghosh and A. Càdenas *Long-time dynamics of complex systems.*, Acc. Chem. Res. **35**, 396-403 (2002).
24. A. Ghosh, R. Elber and H.A. Scheraga *An atomically detailed study of the folding pathways of protein A with the stochastic difference equation*, Proc. Natl. Acad. Sci., U.S.A. **99**, 10394-10398 (2002).
25. M. Khalili, A. Liwo, F. Rakowski, P. Grochowski and H.A. Scheraga *Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode*, J. Phys. Chem. B. **109**, 13785-13797 (2005).
26. M. Khalili, A. Liwo, A. Jagielska and H.A. Scheraga *Molecular dynamics with the united-residue model of polypeptide chains. II. Langevin and Berendsen-bath dynamics and tests on model α-helical systems*, J. Phys. Chem. B **109**, 13798-13810 (2005).
27. A. Liwo, M. Khalili and H.A. Scheraga *Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains*, Proc. Natl. Acad. Sci., U.S.A. **102**, 2362-2367 (2005).
28. M. Khalili, A. Liwo and H.A. Scheraga *Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains*, J.Mol. Biol. **355**, 536-547 (2006).